View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Illinois Digital Environment for Access to Learning and Scholarship Repository

UNIVERSITY OF
ILLINOIS LIBRARY
AT URBANA-CHAMPAIGN
STACKS

Digitized by the Internet Archive
in 2011 with funding from
University of Illinois Urbana-Champaign

http://www.archive.org/details/queueingnetworks646rosb

Faculty Working Papers

QUEUEING NETWORKS UNDER THE CLASS OF STATIONARY SERVICE POLICIES, I.

Zvi Rosberg, Visiting Assistant Professor, Department of Business Administration

#646

College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

FACULTY WORKING PAPERS

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

February 22, 1980

QUEUEING NETWORKS UNDER THE CLASS OF STATIONARY SERVICE POLICIES, I.

Zvi Rosberg, Visiting Assistant Professor, Department of Business Administration

#646

Summary:

We consider an exponential queueing network with different classes of customers, under the class of stationary service policies and the mean cost per unit time as the loss function. It is shown that the queueing network can be transformed, without any loss of generality, into another queueing network in which all the customers at all the service stations have the same service rate. Stability properties of the network under the entire class of service policies are also derived. Moreover, three lower bounds of the loss function under the whole class of service policies, as well as an upper bound of the minimal loss are derived. A convenient form of the loss function is found, which is helpful in finding the form of a good service policy. This form is used in a separate paper, Rosberg [17], to construct a practical method which yields good service policies. Finally, applying the results obtained in this study to a network, where the service requirements are finite mixtures of gamma distributions, is discussed.

1. INTRODUCTION

Queueing models consisting of a network of service stations and different types of customer, moving from one station to another, receiving service at every station are used in many practical applications. The most common use of queueing networks of the type described above is in modeling computer systems, communication networks and combinations of both. An extended review with examples on computer and communications systems modeling by a queueing network is given in Kleinrock [11]. The modeling of a Jobshop-like system (see Jackson [7]) is another practical use of a network of queues. These systems consist of several departments through which different types of jobs move. These jobs must move through several departments before their completion. A multi-stage production line is another possible application of a queueing network. In this application different types of parts of a final product move through several machines and each machine is capable of working on several types of parts.

An extensive use in the last several years of computer networks, networks of microprocessors doing parallel jobs as basic components in the new computer architecture, computerized communication networks and other sophisticated devices have renewed the interest in queueing networks. The studies on queueing networks are mainly concerned with the product form of the stationary distribution of the states of the system under given types of service policies (for recent studies see e.g. Baskett [2], Kelly [10], Chandy [3] and

Noetzel [15]). The insensitivity phenomenon in a product form queueing network relating to the service requirements distribution has also been studied (see e.g. Barbour [1], Schassberger [18] and Jansen [8]), as has the Poisson processes arising from customers streams in the network (Melamed [14]).

The results obtained in the studies mentioned above are applicable to a quite restricted class of service policies. Practical service policies, such as priority rules, are not covered by the models in these studies. Also, the problem of finding an optimal or good service policies did not attain enough interest in a queueing network as in systems with a single service station. For systems with a single service station, optimal service policies have been found under quite general conditions. An optimal rule, which minimizes the number of customers in the queue at any moment of time is given in Schrage [19] for a G/D/1 queue, in which the service duration of a customer is known upon arrival. Optimal service policies in a single service queue with the mean loss per unit of time as the loss function, are given in Cox [5], Kakalik [9] and Klimov [12]. Among these works, Klimov's excellent study is done under the most general conditions. For total discount cost as a loss function, Harrison [6] found the optimal service policy in a M/G/1 queue with different classes of customers. As to systems with more than one service station, almost no work has been carried out yet for finding optimal or good service policies.

In this paper we shall be concerned with an exponential queueing network with different classes of customers under a class of stationary service policies and the mean loss per unit time as the loss function (sections 2.1-2.3). In section 2.4 we shall show how to transform without any loss of generality the queueing network into a simpler one. Some stability properties of the system under the entire class of the service policies will also be derived. Since the problem of finding an optimal service policy in a queueing network is an extremely difficult problem, we are concerned in this paper with finding lower and upper bounds for the loss function under the whole class of service policies (section 3). In section 3 we shall also express the loss function in a way which will be helpful in finding the form of good service policies. This expression of the loss function will be used in a separate paper, Rosberg [17], for constructing a practical method of finding a service policy which reduces the loss function. The method uses simulation and it is programmed for an interactive computer use.

2. THE MODEL AND PRELIMINARY RESULTS

In this section we shall present the queueing network model, the service policies and the Markov processes generated from the model under any given service policy.

2.1. DESCRIPTION OF THE MODEL

We consider an exponential queueing network with different classes of customers and cost for staying in the system, which is defined by the set of parameters

$$\Gamma = (A, B, \lambda, q_\alpha(\beta), \mu_\alpha(\beta), R(\beta), c_\alpha(\beta) \mid \alpha \varepsilon A, \beta \varepsilon B) \text{ , where}$$

A denotes a finite set of service stations $\{1,2,\ldots,a\}$, serving customers independently and simultaneously. Each service station allows an unbounded queue.

B is a finite set $\{1,2,\ldots,b\}$ of classes of customers.

λ is the total arrival rate of customers from outside the system to all the stations. The arrival process is assumed to be Poisson.

$q_\alpha(\beta)$ is the probability that an arriving customer belongs to class β and joins to station α, $\alpha\varepsilon A$, $\beta\varepsilon B$ and $\sum_{\alpha,\beta} q_\alpha(\beta) = 1$.

$\mu_\alpha(\beta)$ is the service rate of customers of class β, $\beta\varepsilon B$, when they are provided service at the service station α, $\alpha\varepsilon A$. The service requirements are exponential r.v.'s, mutually independent and independent of the arrival process.

$R(\beta)$ is a sub-stochastic matrix, which describes the transition probabilities among the service stations of a customer of class β, $\beta\varepsilon B$. The (α,s) element of $R(\beta)$, denoted by $r_{\alpha s}(\beta)$, $\alpha,s\varepsilon A$, is the probability that a customer of class β, $\beta\varepsilon B$, who has been provided service at station α, will move next to station s. The probability that a customer will leave the system is $1 - \sum_s r_{\alpha s}(\beta)$.

$c_\alpha(\beta)$ is the cost of customers of class β, $\beta\varepsilon B$, for staying a unit time at the service station α, $\alpha\varepsilon A$.

To complete the definition of the queueing system, we must still define the service policy, i.e., a decision rule indicating which customer is served at each of the service stations at any moment of time. Let $n = \{n_\alpha(\beta) | \alpha \varepsilon A, \beta \varepsilon B\}$ be the characterization at any moment of time of the queues in system Γ, where $n_\alpha(\beta)$ is the number of customers of class β at service station α at that moment. Furthermore, let any instant of time be a potential decision epoch. We consider any service policy which satisfies the following properties:

(i) At any moment of time the decision rule is a function of the state n only.

(ii) The servers are not allowed to be idle when there are customers at their stations.

(iii) The service of customers at each service station may be interrupted without the loss of any service duration which has already been provided.

Any service policy satisfying (i)-(iii) can be described by a vector function $u(n) = (u_1, u_2, \ldots, u_a)$, where $u_i \varepsilon B \cup \{0\}$ for any $i \varepsilon A$, $u_i = 0$ if and only if $n_i(\beta) = 0$ for any $\beta \varepsilon B$ and $u_i \varepsilon B$ implies $n_i(u_i) > 0$. ($u_i = 0$ means that server i is idle.)

2.2. THE LOSS FUNCTION

In systems operating constantly through time, as computer systems, the mean loss per unit time is usually taken as the loss function. We shall give below the form of this loss function under stationary conditions. Let $n_\alpha^t(\beta)$ be the number of customers of class β at service station α at time t. Also denote

$$n^t = \{n_\alpha^t(\beta)\,|\,\alpha\varepsilon A, \beta\varepsilon B\} \;,\; c = \{c_\alpha(\beta)\,|\,\alpha\varepsilon A, \beta\varepsilon B\} \;,\; (c,n) = \sum_{\alpha,\beta} c_\alpha(\beta) n_\alpha(\beta) \;, \tag{2.1}$$

$$\chi_n(t) = \begin{cases} 1 & n^t = n \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \mathcal{X}_n(T) = \int_0^T \chi_n(t)dt \;.$$

The mean loss per unit time up to time T under any service policy is

$$L_T = E\,\frac{1}{T} \sum_n \mathcal{X}_n(T)(c,n) = \frac{1}{T} \sum_n \int_0^T E\,\chi_n(t)(c,n)dt =$$

$$= \frac{1}{T} \int_0^T \sum_n (c,n) P(n^t=n)dt = \frac{1}{T} \int_0^T (c, En^t)dt \;, \text{ where}$$

$$En^t = \{En_\alpha^t(\beta)\,|\,\alpha\varepsilon A, \beta\varepsilon B\} \;.$$

Under stationary conditions $L_T \xrightarrow[T\to\infty]{} L$, where L is the mean loss per unit time which becomes

$$L = (c, \bar{n}) \;, \tag{2.2}$$

where $\bar{n} = \{\bar{n}_\alpha(\beta)\,|\,\alpha\varepsilon A, \beta\varepsilon B\}$ and $\bar{n}_\alpha(\beta)$ is the expected number of customers of class β at service station α under stationary conditions for a given service policy. The expectation $\bar{n}_\alpha(\beta)$ might also be infinite in which case $L = \infty$. The loss L depends on the service policy chosen and the goal is to minimize L by choosing the proper service policy. Finding an optimal service policy for this model is an extremely difficult problem, which is still open. In this work we find three lower bounds for the loss function and an upper bound for the minimal loss. In a separate paper, Rosberg [17], we shall also give a practical way of finding a service policy which reduces the loss function, as well as a heuristic service policy which is found to work well in some numerical examples. The method is programmed to interactive computer use.

2.3. BASIC ASSUMPTIONS

We make the following basic assumptions about the relative traffic intensity at each service station.

ASSUMPTION 2.3.1. *For any* α, $\alpha\varepsilon A$ *and* β, $\beta\varepsilon B$, *there exists an integer* $k \geq 1$ *such that* $1 - \sum_{s\varepsilon A} r^k_{\alpha s}(\beta) > 0$, *where* $r^k_{\alpha s}(\beta)$ *are the elements of* $R^k(\beta)$, *the* $k^{\underline{th}}$ *power of* $R(\beta)$.

This assumption means that any customer of each class starting service at any station will leave the system with positive probability, after a finite number of visits to the service stations.

For any β, $\beta\varepsilon B$, let $q(\beta) = (q_1(\beta), q_2(\beta), \ldots, q_a(\beta))$ and $\lambda(\beta) = (\lambda_1(\beta), \lambda_2(\beta), \ldots, \lambda_a(\beta))$, $\beta\varepsilon B$ be a solution to

$$\lambda(\beta)(I-R(\beta)) = \lambda q(\beta) . \tag{2.3}$$

Under assumption 2.3.1, $(I-R(\beta))$ is invertible for any $\beta\varepsilon B$. (See, i.e., lemma 3 in Klimov [12].) Together with the nonnegativity of $R(\beta)$ there exists a unique nonnegative solution $\lambda(\beta)$ to equation (2.3).

From (2.3) it follows that $\lambda(\beta) = \lambda q(\beta) \sum_{k=0}^{\infty} R^k(\beta)$, e.g., $\lambda_\alpha(\beta)$, $\alpha\varepsilon A$, is the expected number of potential visits of customers of class β to service station α, arriving to the system from outside, at each unit of time. Let $\rho_\alpha(\beta) = \lambda_\alpha(\beta)/\mu_\alpha(\beta)$. This expression is the potential relative traffic intensity of customers of class β at service station α in the long run.

ASSUMPTION 2.3.2. *For any service station* α, $\alpha\varepsilon A$ $\sum_{\beta\varepsilon B} \rho_\alpha(\beta) < 1$.

2.4. EMBEDDED MARKOV CHAIN

Consider the stochastic process

$$\eta^t = (n^t, i^t) \ , \ t \geq 0 \ ,$$

where n^t is defined as in (2.1) and $i^t = \{i_\alpha^t | \alpha \varepsilon A\}$. Here i_α^t is the class of the customer which is served at station α at time t. For any $\alpha \varepsilon A$, $i_\alpha^t \ \varepsilon \ B\cup\{0\}$, moreover $i_\alpha^t \ \varepsilon \ B$ implies $n_\alpha(i_\alpha^t) > 0$ and $i_\alpha^t = 0$ if and only if $n_\alpha(\beta) = 0$ for any $\beta \varepsilon B$. For definiteness, we assume that $n(0+) = 0$ with probability 1. It is clear that η^t is a Markov process under any given service policy defined in section 2.1.

LEMMA 2.4.1. *Any system Γ can be transformed without any loss of generality into another system with $\mu_\alpha(\beta) = \mu$.*

Proof. We shall change the service rates, $\mu_\alpha(\beta)$ and the transition matrices $R(\beta)$, so as to obtain the same service rate for all customers while leaving the distribution of the process η^t unchanged. The changes are done as follows. Let $\mu = \max\limits_{\alpha \varepsilon A, \beta \varepsilon B} \mu_\alpha(\beta)(1 - r_{\alpha\alpha}(\beta))$ and for any $\alpha \varepsilon A$, $\beta \varepsilon B$ define

$$r^*_{\alpha s}(\beta) = \begin{cases} r_{\alpha s}(\beta)\mu_\alpha(\beta)/\mu & \text{if} \quad s \neq \alpha \ , \\ 1 - (1 - r_{\alpha\alpha}(\beta))\mu_\alpha(\beta)/\mu & \text{if} \quad s = \alpha \ . \end{cases}$$

Now, let the redefined system be $\Gamma^* = (A, B, \lambda, q_\alpha(\beta), \mu, R^*(\beta), c_\alpha(\beta) | \alpha \varepsilon A, \beta \varepsilon B)$. For any given service policy, let η^t and η^{*t} be the Markov processes generated from the same service policy by the system Γ and Γ^* respectively. From the exponential properties of the processes η^t and η^{*t}, it is easy to see that the Markov processes of the pure jumps embedded in those processes are the same. Furthermore, the processes of the time lengths between pure jumps in η^t and η^{*t} are also the same. (A pure jump in the process η^t is a transition accompanied with a change in the state of the process and the pure jumps process embedded in η^t is the process defined by η^t at the moments of time when pure jumps occur.) (e.g., see Lippman [13] or Cinler [4], Chap. 8.) Thus, under any service policy, η^t and η^{*t} have the same distribution. This completes the proof.

From now on we shall consider only systems Γ^* with $\mu_\alpha(\beta) = \mu$ for any $\alpha \varepsilon A$, $\beta \varepsilon B$ and shall omit the star in our notation. In addition to this reformulation we add to each service station one dummy customer, denoted by 0, who is served when and only when, there are no other regular customers at the service station. The parameters of the dummy customers in the system Γ are $c_\alpha(0) = 0$, $q_\alpha(0) = 0$, $\mu_\alpha(0) = \mu$ and $r_{\alpha\alpha}(0) = 1$ for any $\alpha \varepsilon A$. Let $B^* = B \cup \{0\}$, where B is the set of classes of regular customers. The reformulation proposed by lemma 2.4.1 and the addition of dummy customers leave the probabilistic behaviour of η^t and the loss function unchanged while at the same time greatly simplify the analysis.

We shall observe the process η^t at the moments of time when transitions occur. (Not necessarily caused by a change in the state.) A transition occurs if either one of the following occurs:

(i) A new customer from outside arrives at one of the service station.

(ii) One of the customers, possibly the dummy one, finishes being served at one of the service stations. (The customer might immediately return to the same station in which case

Let N, N = 1,2,3,... denote the sequence of transitions and let the time period between the N-1 and the N transitions be called the $N^{\underline{th}}$ step. For any $\alpha\varepsilon A$, $\beta\varepsilon B$ let $n_\alpha^N(\beta)$ be the number of customers of class β at station α *immediately after* the $N^{\underline{th}}$ transition has occurred and let $n^N = \{n_\alpha^N(\beta)|\alpha\varepsilon A,\ \beta\varepsilon B\}$. Also, let i_α^N denote the class of the customer provided service at station α *immediately before* the $N^{\underline{th}}$ transition occurs and let $i^N = \{i_\alpha^N|\alpha\varepsilon A\}$. To be consistent with the description of a service policy in section 2.1 let $u(n^N) = (u_1,u_2,\ldots,u_a)$ be the vector of classes of customers which are provided service at each station *immediately after* the $N^{\underline{th}}$ transition has occurred.

The process $\eta^N = (n^N,i^N)$, N = 1,2,... is an embedded Markov chain in the process η^t, $t \geq 0$, having a denumerable number of states. The stationary distribution of η^N (if it exists) is the same as of η^t, since all the pure jumps of η^t are covered by η^N and the lengths of time between any two transitions in η^N are equally distributed. Therefore, it is sufficient to analyze the embedded Markov chain η^N.

LEMMA 2.4.2. *For any given state of the Markov chain η^N under any given service policy, the probability that the next transition will occur due to*

(a) the completion of service by a customer (including a dummy one) at the service station α, is μ/Λ, for any $\alpha\varepsilon A$.

(b) an arrival of a customer of class β to station α from outside, is $\lambda q_\alpha(\beta)/\Lambda$, for any $\alpha\varepsilon A$, $\beta\varepsilon B$, where $\Lambda = \lambda + a\mu$.

Proof. The lemma follows immediately from the fact that the service requirements and the interarrival lengths of time between two customers are exponentially distributed. This follows by direct computation of the probabilities in which the reformulation of Γ is crucial.

For any $\alpha \varepsilon A$, let $S_0^\alpha = \{n | n_\alpha(\beta) = 0$ for any $\beta \varepsilon B\}$. Whenever $n^t \varepsilon S_0^\alpha$, there is only a dummy customer at station α. Also, let τ_α denote the first re-entry time to S_0^α and $E_{(n,i)}(\tau_\alpha)$ the expectation of τ_α, given that $\eta^1 = (n,i)$.

LEMMA 2.4.3. *Under assumptions 2.3.1 and 2.3.2, the Markov chain* η^N, $N = 1,2,\ldots$, *under any service policy, is irreducible, aperiodic and satisfies the following properties:*

(a) *For any* α, $\alpha \varepsilon A$ and state (n,i), $E_{(n,i)}(\tau_\alpha) \leq B_\alpha(n) < \infty$, *where* $B_\alpha(n)$ *is a number independent of the service policy.*

(b) *For any* α, $\alpha \varepsilon A$ *and state* (n,i)

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} P(\eta^j \varepsilon S_0^\alpha / \eta^1 = (n,i)) \geq 1 - \sum_{\beta \varepsilon B} \rho_\alpha(\beta) > 0 .$$

Proof. The irreducibility and aperiodicity are easily checked from the transition probabilities of η^N. Part (a) of the lemma is proved in Rosberg [16], section 3, and the bounds $B_\alpha(n)$ are given explicitly. The proof of part (b) follows along the same lines as the proof of theorem 2 in Rosberg [16] using the results given in section 3 there.

REMARKS

(i) The bounds $B_\alpha(n)$, which are explicitly given in Rosberg [16], are easily computed and may be used in practice for evaluating the performance of a networks of queues under any service policy. As an example, $B_\alpha(0) = 1/(1 - \sum_{\beta\varepsilon B} \rho_\alpha(\beta))$, where 0 stands for the zero matrix.

(ii) Part (b) of lemma 2.4.3 means that the long run proportion of time that any service station α, $\alpha\varepsilon A$, stays empty is positive. That is, under assumptions 2.3.1 and 2.3.2, the network under any service policy, is not overloaded and the number of customers at each service station does not explode to infinity. This property is weaker than egrodicity and ensures the stability of the network.

(iii) Assumptions 2.3.1 and 2.3.2 are sufficient for η^N to be ergodic under certain service policies (for example the Serive-Sharing policy), but it is still an open question whether these assumptions are sufficient for ergodicity under an arbitrary service policy.

In the analysis in the next section we shall consider only service policies under which η^N is ergodic and the loss function L is finite. Other service policies are not of interest since they provide an infinite loss.

3. BOUNDS OF THE LOSS FUNCTION

In this section we shall represent the loss function, L, by a sum of two expressions. The first expression, which is independent of the service policy, is used as a lower bound for L. The second one is used to construct a subclass of service policies which reduce the loss toward the lower bound. This rerresentation is derived by using the technique used in Klimov [12] for a single server station.

In addition we find two other lower bounds and one upper bound. These bounds are used to evaluate the service policies in the class suggested above.

3.1. THE PROCESS η^N

Let η^N be a Markov chain generated by Γ and any given service policy. For any set $n = \{n_\alpha(\beta) | \alpha \varepsilon A, \beta \varepsilon B^*\}$ and any set $Z = \{Z_\alpha(\beta) | \alpha \varepsilon A, \beta \varepsilon B^*\}$ of numbers $Z_\alpha(\beta)$, let

$$Z^n = \prod_{\substack{\alpha \varepsilon A \\ \beta \varepsilon B^*}} Z_\alpha(\beta)^{n_\alpha(\beta)} .$$

Further, for any set $\underline{\beta} = \{\beta_\alpha | \alpha \varepsilon A, \beta_\alpha \varepsilon B^*\}$ and any set $i^N = \{i_\alpha^N | \alpha \varepsilon A\}$, let

$$\delta_{\underline{\beta}, i^n} = \prod_{\alpha \varepsilon A} \delta_{\beta_\alpha, i_\alpha^N} ,$$

where $\delta_{\beta_\alpha, i_\alpha^N}$ is the Kronecker delta. We shall write $0 \leq Z \leq 1$ iff $0 \leq Z_\alpha(\beta) \leq 1$ for all $\alpha \varepsilon A$, $\beta \varepsilon B^*$ and $Z = 1$ if $Z_\alpha(\beta) = 1$ for all $\alpha \varepsilon A$, $\beta \varepsilon B^*$. For $0 \leq Z \leq 1$ and any $\underline{\beta}$, let

$$P_{\underline{\beta},N}(Z) = E(Z^{n^N} \delta_{\underline{\beta},i^N}) \; ; \; R_{\underline{\beta},N}(Z) = E(Z^{n^N} \delta_{\underline{\beta},u(n^N)})$$

$$P_N(Z) = E(Z^{n^N}) .$$

We have, $P_N(Z) = \sum_{\underline{\beta}} P_{\underline{\beta},N}(Z) = \sum_{\underline{\beta}} R_{\underline{\beta},N}(Z)$.

For $Z_\alpha(\beta) = 0$ let $R_{\underline{\beta},N}(Z)/Z_\alpha(\beta) \triangleq E(Z^{n'^N} \delta_{\underline{\beta},i^N})$, where n'^N is derived from n^N by removing one customer of type β from station α.

LEMMA 3.1.1. *For any possible $\underline{\beta}$ we have*

$$P_{\underline{\beta},N+1}(Z) = \sum_{\alpha\varepsilon A}\left\{\frac{R_{\underline{\beta},N}(Z)}{Z_\alpha(\beta_\alpha)}\ \frac{\mu}{\Lambda}\ Q_\alpha^{\beta_\alpha}(Z)\right\} + R_{\underline{\beta},N}(Z)\ \frac{\lambda}{\Lambda} \sum_{\substack{\alpha\varepsilon A\\ \beta\varepsilon B}} q_\alpha(\beta)Z_\alpha(\beta)\ ,$$

where

$$Q_\alpha^{\beta_\alpha}(Z) = \sum_{s\varepsilon A}\left(r_{\alpha s}(\beta_\alpha)Z_s(\beta_\alpha)\right) + \left(1 - \sum_{s\varepsilon A} r_{\alpha s}(\beta_\alpha)\right)\ .$$

Proof. For any $\alpha\varepsilon A$, $\beta\varepsilon B^*$ interpret the number $Z_\alpha(\beta)$ as the probability that a customer of class β entering service station α (after some service in any service station or upon entering the system), is "colored by red" and $1 - Z_\alpha(\beta)$ as the probability that he is "colored by blue." Thus, customers may change their colors after any demand for service. The color assignments are mutually independent and independent of the service requirements and the arrival process. Now, with this probabilistic interpretation we have the following:

> $P_{\underline{\beta},N+1}(Z)$ is the probability that at the N+1 step a customer of class β_α was served at station α, for all $\alpha\varepsilon A$ and after this step was completed no blue customer remained in the system.
>
> $R_{\underline{\beta},N}(Z)$ is the probability that after N steps are completed, all the remaining customers are red and at the next step a customer of class β_α is served at station α, for all $\alpha\varepsilon A$.
>
> $R_{\underline{\beta},N}(Z)/Z_\alpha(\beta_\alpha)$ is the probability that after N steps are completed, all the remaining customers are red excluding possibly the customer which will

the next step a customer of class β_i is served at station i for all $i\varepsilon A$.

$q_\alpha(\beta)Z_\alpha(\beta)$ is the probability that an arriving customer belongs to class β, joins to station α and is colored by red.

$Q_\alpha^{\beta_\alpha}(Z)$ is the probability that a customer of class β_α which has been served at station α does not become blue after being served.

The lemma now follows from the probabilistic interpretations above and from lemma 2.4.2.

LEMMA 3.1.2. *The following limits exist*

$$\lim_{N\to\infty} P_{\underline{\beta},N}(Z) = P_{\underline{\beta}}(Z) , \qquad \lim_{N\to\infty} R_{\underline{\beta},N}(Z) = R_{\underline{\beta}}(Z) ,$$

$$\lim_{N\to\infty} P_N(Z) = P(Z) .$$

In addition,

$$P_{\underline{\beta}}(Z) = \sum_{\alpha\varepsilon A}\left(\frac{R_{\underline{\beta}}(Z)}{Z_\alpha(\beta_\alpha)}\frac{\mu}{\Lambda}Q_\alpha^{\beta_\alpha}(Z)\right) + R_{\underline{\beta}}(Z)\frac{\lambda}{\Lambda}\sum_{\substack{\alpha\varepsilon A\\ \beta\varepsilon B}} q_\alpha(\beta)Z_\alpha(\beta) , \tag{3.1}$$

and under any service policy

$$\lim_{N\to\infty} P(i_\alpha^N = \beta) = \sum_{\underline{\beta}} R_{\underline{\beta}}(1)\delta_{\beta_\alpha,\beta} = \sum_{\underline{\beta}} P_{\underline{\beta}}(1)\delta_{\beta_\alpha,\beta} = \rho_\alpha(\beta) \text{ for any } \alpha\varepsilon A,\ \beta\varepsilon B ,$$

$$\lim_{N\to\infty} P(i_\alpha^N = 0) = \sum_{\underline{\beta}} R_{\underline{\beta}}(1)\delta_{\beta_\alpha,0} = \sum_{\underline{\beta}} P_{\underline{\beta}}(1)\delta_{\beta_\alpha,0} = 1 - \sum_{\beta\varepsilon B}\rho_\alpha(\beta) > 0 \text{ for any } \alpha\varepsilon A . \tag{3.2}$$

Proof. From lemma 2.4.3, η^N is irreducible and aperiodic and from our assumption it is also ergodic. Thus, all the limits in the lemma exist and (3.1) follows directly from lemma 3.1.1. For any $i\varepsilon A$ and $j\varepsilon B^*$ let,

$$x_{i,j} = \frac{\partial}{\partial Z_i(j)} P(Z)\Big|_{Z=1} \tag{3.3}$$

and, for any possible $\underline{\beta}$, let,

$$x_{i,j}(\underline{\beta}) = \frac{\partial}{\partial Z_i(j)} R_{\underline{\beta}}(Z)\Big|_{Z=1} . \tag{3.4}$$

From the definitions of $P(Z)$ and $R_{\underline{\beta}}(Z)$, it follows that

$$x_{i,j} = \sum_{\underline{\beta}} x_{i,j}(\underline{\beta}) . \tag{3.5}$$

From (3.1) and the definitions of $P(Z)$ and $R_{\underline{\beta}}(Z)$ we have

$$P(Z) = \sum_{\underline{\beta}} P_{\underline{\beta}}(Z) = \sum_{\underline{\beta}} R_{\underline{\beta}}(Z)\{\frac{\lambda}{\Lambda} \sum_{\alpha,\beta} q_\alpha(\beta) Z_\alpha(\beta) + \frac{\mu}{\Lambda} \sum_\alpha Q_\alpha^{\beta_\alpha}(Z)/Z_\alpha(\beta_\alpha)\} . \tag{3.6}$$

Taking the partial derivatives with respect to $Z_i(j)$, $i\varepsilon A$, $j\varepsilon B$, on the left and the right hands of (3.6) at the point $Z = 1$, we obtain from the definitions in (3.3) and (3.4), that for any $i\varepsilon A$ and $j\varepsilon B$

$$x_{i,j} = \sum_{\underline{\beta}} \{x_{i,j}(\underline{\beta}) + R_{\underline{\beta}}(1)[\frac{\lambda}{\Lambda} q_i(j) + \sum_\alpha \frac{\mu}{\Lambda} r_{\alpha,i}(j)\delta_{\beta_\alpha,j} - \frac{\mu}{\Lambda} \delta_{\beta_i,j}]\} . \tag{3.7}$$

From ergodicity it follows that $\sum_{\underline{\beta}} R_{\underline{\beta}}(1) = 1$. Thus, from (3.5) and (3.7) we obtain that for any $i\varepsilon A$ and $j\varepsilon B$

$$\sum_{\underline{\beta}} \mu R_{\underline{\beta}}(1)\delta_{\beta_i,j} = \lambda q_i(j) + \sum_{\alpha\varepsilon A} r_{\alpha,i}(j) \sum_{\underline{\beta}} \mu R_{\underline{\beta}}(1)\delta_{\beta_\alpha,j} . \tag{3.8}$$

For any $\alpha\varepsilon A$ and $\beta\varepsilon B$ let $y_\alpha(\beta) = \lim_{N\to\infty} P(i_\alpha^N = \beta)$. Hence, $y_\alpha(\beta) = \sum_{\underline{\beta}} R_{\underline{\beta}}(1)\delta_{\beta_\alpha,\beta}$. Using this expression for $y_\alpha(\beta)$ in equation (3.8) we have, for any $\alpha\varepsilon A$ and $\beta\varepsilon B$,

$$\mu\, y_\alpha(\beta) = \lambda\, q_\alpha(\beta) + \sum_{s\varepsilon A} r_{s,\alpha}(\beta)\mu y_s(\beta) . \tag{3.9}$$

From section 2.3 and assumptions 2.3.1 and 2.3.2 we have

$$y_\alpha(\beta) = \rho_\alpha(\beta) \text{ , for any } \alpha\varepsilon A,\ \beta\varepsilon B \qquad (3.10)$$

and from ergodicity,

$$y_\alpha(0) = 1 - \sum_{\beta\varepsilon B} \rho_\alpha(\beta) > 0 \text{ for any } \alpha\varepsilon A \ . \qquad (3.11)$$

From (3.1) it follows that

$$R_{\underline{\beta}}(1) = P_{\underline{\beta}}(1) \quad \text{for any possible } \underline{\beta} \ . \qquad (3.12)$$

Equations (3.10)-(3.12) complete the proof.

One may define a loss function L', as the mean loss per unit time arising only from the waiting time of the customers (i.e., excluding the losses due to time being served). In the same way as used in section 2.2 we obtain that $L' = (c,\overline{\ell})$, where $\overline{\ell} = \{\ell_\alpha(\beta)|\alpha\varepsilon A,\ \beta\varepsilon B\}$ and $\overline{\ell}_\alpha(\beta)$ is the expected number of customers of class β in the queue of station α (excluding the one who is possibly being served). From (3.2), it is easy to see that under any service policy

$$\overline{n}_\alpha(\beta) = \overline{\ell}_\alpha(\beta) + \rho_\alpha(\beta) \text{ , thus}$$

$$L = L' + \sum_{\alpha,\beta} c_\alpha(\beta)\rho_\alpha(\beta) \ . \qquad (3.13)$$

From (3.13) we obtain the following corollary.

COROLLARY 3.1.1. The preference order among the service policies is the same under the loss functions L and L', where the preference order is defined according to the values of the loss functions under the set of service policies.

3.2. A LINEAR PROGRAMMING PROBLEM

The problem of finding a service policy which minimizes the loss function is a dynamic programming problem with a denumerable state space. We shall transform this problem into a linear programming problem with a finite number of variables. The linear programming problem will have a larger space of feasible solutions than the space of feasible service policies. For any $i,k \in A$ and $j,m \in B^*$ let

$$y_{i,j}(k,m) = \sum_{\underline{\beta}} x_{i,j}(\underline{\beta})\delta_{\beta_k,m} .$$

The variable $y_{i,j}(k,m)$, $i,k \in A$, $j,m \in B^*$, is the expected number of customers of class j at station i, under stationary conditions, given that a customer of class m is provided service at station k, times the probability that a customer of class m is provided service at station k. The variables $y_{i,j}(k,m)$ are closely related to the service policies and are convenient for optimization purpose.

LEMMA 3.2.1. *For any Markov chain* η^N, $i,k \in A$ *and* $j,m \in B^*$ *we have*

$$\begin{aligned} y_{i,j}(k,m) + y_{k,m}(i,j) &= \sum_{\alpha \in A} r_{\alpha,k}(m) y_{i,j}(\alpha,m) + \sum_{\alpha \in A} r_{\alpha,i}(j) y_{k,m}(\alpha,j) + \\ &+ \frac{\lambda}{\mu} q_k(m) \sum_{\beta \in B} y_{i,j}(i,\beta) + \frac{\lambda}{\mu} q_i(j) \sum_{\beta \in B} y_{k,m}(k,\beta) + t(i,j,k,m) , \end{aligned} \tag{3.14}$$

where

$$t(i,j,k,m) = \begin{cases} 2\rho_i(j)(1-r_{ii}(j)) & \text{if } i=k \text{ and } j=m , \\ -(\rho_i(m) r_{i,k}(m) + \rho_k(m) r_{k,i}(m)) & \text{if } i \neq k \text{ and } j=m , \\ 0 & \text{otherwise} . \end{cases} \tag{3.15}$$

Proof. The lemma follows by equating the partial derivatives of both sides of formula (3.6) with respect to $Z_i(j)$ and $Z_k(m)$ at the point $Z = 1$. Indeed,

$$\frac{\partial}{\partial Z_i(j)} R_{\underline{\beta}}(Z)\Big|_{Z=1} = x_{i,j}(\underline{\beta}) ,$$

$$\frac{\partial}{\partial Z_i(j)} \frac{R_{\underline{\beta}}(Z)}{Z_\alpha(\beta_\alpha)}\Bigg|_{Z=1} = x_{i,j}(\underline{\beta}) - R_{\underline{\beta}}(1)\delta_{(i,j),(\alpha,\beta_\alpha)} ,$$

$$\frac{\partial^2}{\partial Z_i(j)\partial Z_k(m)} \frac{R_{\underline{\beta}}(Z)}{Z_\alpha(\beta_\alpha)}\Bigg|_{Z=1} = \frac{\partial^2}{\partial Z_i(j)\partial Z_k(m)} R_{\underline{\beta}}(Z)\Big|_{Z=1} - x_{i,j}(\underline{\beta})\delta_{(\alpha,\beta_\alpha),(k,m)} -$$

$$- x_{k,m}(\underline{\beta})\delta_{(\alpha,\beta_\alpha),(i,j)} + 2R_{\underline{\beta}}(1)\delta_{(\alpha,\beta_\alpha),(i,j)}\delta_{(\alpha,\beta_\alpha),(k,m)} ,$$

$$\frac{\partial}{\partial Z_i(j)}\left(\frac{\lambda}{\Lambda}\sum_{\alpha,\beta} q_\alpha(\beta)Z_\alpha(\beta)\right)\Bigg|_{Z=1} = \frac{\lambda}{\Lambda} q_i(j) ,$$

$$\frac{\partial^2}{\partial Z_i(j)\partial Z_k(m)}\left(\frac{\lambda}{\Lambda}\sum_{\alpha,\beta} q_\alpha(\beta)Z_\alpha(\beta)\right)\Bigg|_{Z=1} = 0 ,$$

$$\frac{\partial}{\partial Z_i(j)} Q_\alpha^{\beta_\alpha}(Z)\Big|_{Z=1} = r_{\alpha,i}(\beta_\alpha)\delta_{\beta_\alpha,j} ,$$

$$\frac{\partial^2}{\partial Z_i(j)\partial Z_k(m)} Q_\alpha^{\beta_\alpha}(Z)\Big|_{Z=1} = 0 ,$$

$$\frac{\partial^2}{\partial Z_i(j)\partial Z_k(m)} P(Z)\Big|_{Z=1} = \frac{\partial^2}{\partial Z_i(j)\partial Z_k(m)}\left(\sum_{\underline{\beta}} R_{\underline{\beta}}(Z)\right)\Bigg|_{Z=1} = \sum_{\underline{\beta}} \frac{\partial^2}{\partial Z_i(j)\partial Z_k(m)} R_{\underline{\beta}}(Z)\Big|_{Z=1} .$$

In addition, from the definition of $y_{i,j}(k,m)$ we have for any $i\varepsilon A$ and $j\varepsilon B$

$$y_{i,j}(i,0) = 0 ,$$

and

$$x_{i,j} = \sum_{\beta\varepsilon B^*} y_{i,j}(i,\beta) = \sum_{\beta\varepsilon B} y_{i,j}(i,\beta) \; .$$

Now, the lemma follows by direct computation from (3.6) and the above formulas.

From the definitions of $x_{\alpha,\beta}$ and $y_{\alpha,\beta}(k,m)$ it follows for any $\alpha\varepsilon A$ and $\beta\varepsilon B$,

$$\bar{n}_\alpha(\beta) = x_{\alpha,\beta} = \sum_{m\varepsilon B^*} y_{\alpha,\beta}(k,m) \; , \text{ for any } k\varepsilon A \tag{3.16}$$

and

$$y_{\alpha,\beta}(k,0) = \sum_{m\varepsilon B} y_{\alpha,\beta}(\alpha,m) - \sum_{m\varepsilon B} y_{\alpha,\beta}(k,m) \; , \text{ for any } k\varepsilon A \; . \tag{3.17}$$

From (2.2) and (3.16) we have,

$$L = \frac{1}{a} \sum_{\alpha,\beta} c_\alpha(\beta) \sum_{\substack{k\varepsilon A \\ m\varepsilon B^*}} y_{\alpha,\beta}(k,m) \; . \tag{3.18}$$

The variables $y_{\alpha,\beta}(k,m)$, $\alpha,k\varepsilon A$, $\beta,m\varepsilon B^*$ must satisfy the inequalities (among possibly other constraints),

$$y_{\alpha,\beta}(k,m) \geq 0 \text{ for any } \alpha,k\varepsilon A,\; \beta,m\varepsilon B^* \; , \tag{3.19}$$

and the equations (3.14) and (3.17). Minimizing L, given in (3.18), subject to the linear constraints (3.14), (3.17) and (3.19) is a finite linear programming problem.

To simplify the problem, we shall represent it in a matrix form which has a special structure. Define the matrices $Y = (y(i,j))$ and $T = (t(i,j))$, the column vectors $q = (q(i))$ and $c = (c(i))$ as follows:

$$y((m-1)a+k,(j-1)a+i)) = \begin{cases} y_{i,j}(k,m) & \text{if } 1\le i,k\le a \text{ and } 1\le j,m\le b \text{ ,} \\ y_{i,j}(k,0) & \text{if } 1\le i,k\le a; 1\le j\le b \text{ and } m=b+1 \text{ ,} \\ 0 & \text{if } 1\le i,k\le a; 1\le m\le b+1 \text{ and } j=b+1 \text{ .} \end{cases}$$

The element $y((m-1)a+k,(j-1)a+i)$ is the element of the $(m-1)a+k$ row and the $(j-1)a+i$ column. The matrix Y so defined has $(b+1)a$ rows and columns.

$$t((j-1)a+i,(m-1)a+k) = \begin{cases} t(i,j,k,m) & \text{if } 1\le i,k\le a \text{ and } 1\le j,m\le b \text{ ,} \\ 0 & \text{if } 1\le i,k\le a; 1\le j\le b \text{ and } m=b+1 \text{ ,} \\ 0 & \text{if } 1\le i,k\le a; 1\le m\le b+1 \text{ and } j=b+1 \text{ ,} \end{cases}$$

where $t(i,j,k,m)$ are the elements defined in (3.15). The element $t((j-1)a+i,\ (m-1)a+k)$ is the element of the $(j-1)a+i$ row and the $(m-1)a+k$ column. The matrix T, so defined has $(b+1)a$ rows and columns.

$$c((m-1)a+k) = \begin{cases} c_k(m) & \text{if } 1 \le k \le a \text{ and } 1 \le m \le b \text{ ,} \\ 0 & \text{if } 1 \le k \le a \text{ and } m = b + 1 \text{ .} \end{cases}$$

$$q((m-1)a+k) = \begin{cases} q_k(m) & \text{if } 1 \le k \le a \text{ and } 1 \le m \le b \text{ ,} \\ 0 & \text{if } 1 \le k \le a \text{ and } m = b + 1 \text{ .} \end{cases}$$

Also, define the matrix

$$R = \begin{bmatrix} R(1) & & & 0 \\ & R(2) & & \\ & & \ddots & \\ 0 & & & R(b) \\ & & & & I \end{bmatrix},$$

where I is the identity matrix of rank a. Let $\underline{1} = (1,1,\ldots,1)'$ be a $(b+1)a$ element column vector, all whose elements are one.

Using the notations above we have that $L = (c, Y'\underline{1})/a$. Now minimizing L is the same as minimizing $(c, Y'\underline{1})$. Thus, the linear programming problem given in (3.14), (3.17), (3.18) and (3.19) becomes

$$\min_{Y} \, (c, Y'\underline{1}) \;, \tag{3.20}$$

where the element $Y = (y(i,j))$ satisfy

$$Y \geq 0$$

and (3.21)

$$Y + Y' = Y'R + R'Y + \frac{\lambda}{a\mu} Y'\underline{1}q' + \frac{\lambda}{a\mu} q1'Y + T \;,$$

where X' is the transpose of the matrix X.

REMARK 3.2.1.

(i) Any feasible service policy corresponds to a matrix Y satisfying the set of constraints (3.21), but not vise versa. This is clear from the way we constructed the linear programming problem. Hence, solving the linear programming problem (3.20)-(3.21), does not yet provide the optimal service policy.

(ii) The set of constraints (3.21) does not include the constraints given in (3.17) in their full strength. The reason is that we want to obtain the convenient structure of the constraints as given in (3.21). The constraints in (3.17) will be used later to obtain a better lower bound for L.

3.3. ANOTHER FORM OF THE LOSS FUNCTION L

We use the special structure of the linear programming problem given in (3.20) and (3.21) to obtain a form of the loss function, for which the constraints are already built-in. The loss function so obtained, provide a lower bound for L and gives an idea how to construct good service policies.

Since the algebraic structure of our linear programming problem has the same algebraic structure as the linear programming problem in Klimov [12], sections 9, 10 (but representing a completely different problem), we shall use the notation and the results appearing there.

Let $\Omega = \{1,2,\ldots,(b+1)a\}$ be a set of phases of service. For any i, $1\leq i\leq a$ and j, $1\leq j\leq b$, the element $(j-1)a+i$ in Ω corresponds to a phase of service of a customer of class j at station i. The element ba+i corresponds to a phase of service of a dummy customer at station i. Note that, any phase $p\varepsilon\Omega$, determined uniquely the class of the customer and the service station. We shall define a complete order, henceforth "Klimov order," among the phases of service in Ω, as following.

Let $\gamma = (\gamma_1,\gamma_2,\ldots,\gamma_{(b+1)a})'$ denote a column vector and let M be any subset of Ω. Denote by $\gamma(M)$ the vector obtained from γ after removing the elements γ_i, for all $i\varepsilon\Omega\setminus M$. Further, let R(M) be the matrix obtained from R after removing all the columns and rows with indexes in $\Omega\setminus M$.

Since $(I-R(\beta))$ is invertible for every $\beta\varepsilon B$, so is $(I-R)$. Thus, for any $M\subseteq\Omega$, $(I(M)-R(M))$ is invertible, where $I(M)$ is the identity matrix of order $\#M$.

Let $\gamma(M) = (\gamma_i(M))$, $M\subseteq\Omega$, be the solution of the equation

$$(I(M) - R(M))\gamma(M) = 1(M) , \tag{3.22}$$

where $1(M) = (1,1,\ldots,1)'$ is a column vector of order $\#M$.

The element $\gamma_i(M)$, $i\varepsilon M$ of the vector $\gamma(M)$, is the mean total number of visits to phases of service in M, of a customer starting with phase i up to its first exit from M.

Define a sequence of phases, $p_{a(b+1)},\ldots,p_2,p_1$ by the following recursive relations:

$$M_{a(b+1)} = \Omega , \quad c_p(M_{a(b+1)}) = c(p) \quad \text{for any} \quad p\varepsilon\Omega ;$$

$$\begin{aligned} \frac{c_{p_i}(M_i)}{\gamma_{p_i}(M_i)} &= \min_{p\varepsilon M_i} \frac{c_p(M_i)}{\gamma_p(M_i)} = m_i , \quad p_i\varepsilon M_i ; \\ M_{i-1} &= M_i\setminus\{p_i\} \end{aligned} \tag{3.23}$$

$$c_p(M_{i-1}) = \gamma_p(M_i)\left[\frac{c_p(M_i)}{\gamma_p(M_i)} - m_i\right] \geq 0 , \quad \text{for any} \quad p\varepsilon M_{i-1} .$$

REMARK 3.3.1.

(i) It is clear that the ordered sequence $p_{a(b+1)},\ldots,p_2,p_1$ need not be uniquely determined since m_i in (3.23) may be attained for several phases. Any order of the sequence is suitable.

(ii) The set of phases $\{p_{ba+i} | 1 \le i \le a\}$ can be obtained by the procedure given in (3.23), in any order among themself and will always correspond to the phases of service of a dummy customer at station i, for all $i \varepsilon A$. For convenience of the notation we shall choose the order such that phase p_{ba+i}, $i \varepsilon A$, correspond to the phase of service of a dummy customer in station i.

DEFINITION 3.3.1. *For any* $i,j \varepsilon \Omega$, *the phase* p_i *precedes phase* $p_j (p_i \prec p_j)$ *according to the "Klimov order" if* $i < j$.

Now, rename the set of phases $\Omega = \{p_1, p_2, \ldots, p_{(b+1)a}\}$ such that, phase $p_i \varepsilon \Omega$ will be denoted by i. Note that under this renaming, for any $p, p' \varepsilon \Omega$, $p \prec p'$ if and only if $p < p'$.

With this new notation we shall change the elements in Y, R, T, q and c accordingly. For any $i \varepsilon \Omega$, let the $i^{\underline{th}}$ row of Y, R, T, q and c be replaced by its $p_i^{\underline{th}}$ row and similarly for the columns of Y, R and T.

Henceforth the notations $\Omega = \{1,2,\ldots,(b+1)a\}$, Y, R, T, q and c will be used to denote the appropriate reordered phases, matrices and vectors. It is clear that the problem and the results obtained so far are not affected by this renaming.

For any p, $1 \le p \le (b+1)a$, let M_p, $\gamma_i(M_p)$, $c_i(M_p)$, $i \le p$, be the elements obtained by the procedure given in (3.23). For any p, $1 \le p \le ba$, define the following vectors, each with $(b+1)a$ elements:

$$u_p = (\gamma_1(M_p), \gamma_2(M_p), \ldots, \gamma_p(M_p), 0, \ldots, 0)'$$

and (3.24)

$$v_p = (0,0,\ldots,0,1,1,\ldots,1)' ,$$

where v_p has p zeros preceding the one's.

Further, let

$$z_p = c_p(M_p)/\gamma_p(M_p) ,$$

$$h_p = (v_p, Yu_p)/(1-\lambda(q,u_p)/a\mu) ,$$

$$g_p = (Tu_p, u_p)/2(1-\lambda(q,u_p)/a\mu) , \qquad (3.25)$$

$$z = (z_1, z_2, \ldots, z_{ab})' , \quad h = (h_1, h_2, \ldots, h_{ab})'$$

and
$$g = (g_1, g_2, \ldots, g_{ab})' .$$

THEOREM 3.3.1. *Under any service policy*

$$L = (z,g)/a + (z,h)/a .$$

In addition, $z \geq 0$ *and* $h \geq 0$.

Proof. From (3.18) we have

$$L = (c, Y'1)/a . \qquad (3.26)$$

The linear programming problem in (3.20) and (3.21) has the same algebraic structure as the linear programming problem in Klimov [12], section 9. Following the same procedure as in Klimov [12], sections 9, 10, we obtain the theorem from (3.26) and the definitions given in (3.24) and (3.25).

REMARK 3.3.2.

Theorem 3.3.1 gives a representation of L, as a sum of two expressions. The first one is independent of the service policy and is used as a lower bound of L. The second one is dependent on the service policy through Y and is used in a separate paper, Rosberg [17], for constructing service policies which reduce the loss function.

3.4. BOUNDS OF THE LOSS FUNCTION

The theorems below contain three lower bounds of the loss function, none of which is an infimum. Also, an upper bound of the minimum loss is derived. The bounds have importance for evaluation of systems Γ and service policies.

THEOREM 3.4.1. *Under any service policy*

$$L \geq (z,g)/a \triangleq LB1 \ .$$

Proof. The theorem follows directly from theorem 3.3.1 and remark 3.3.2.

REMARK 3.4.1.

From the definitions in (3.24) and (3.25) it can be shown (see Rosberg [17]) that

$$(z,h) = \sum_{p'=1}^{ab} \sum_{p=p'+1}^{a(b+1)} w(p,p')y(p,p') \ , \tag{3.27}$$

where, $w(p,p')$ are nonnegative constants independent of the service policy, $y(p,p')$ are the elements of Y, whose rows and columns are ordered according to "Klimov order."

By recalling the form of the linear programming problem (3.20) and (3.21), it also can be shown that, there exists a matrix Y such that $y(p,p') = 0$ for any p', $1 \leq p' \leq ab$ and $p, p'+1 \leq p \leq a(b+1)$

which also satisfies (3.21). Thus, from theorem 3.3.1 and (3.27) it follows that LB1 is the best lower bound which can be derived from the linear programming problem (3.20)-(3.21). Note that LB1 can be computed without solving the linear programming problem.

From remark 3.2.1 (ii) it is clear that we can obtain a better lower bound than LB1, by solving the linear programming problem (3.20)-(3.21) with the additional constraints given in (3.17). In this case we cannot find the minimum value of the objective function given in (3.20), without solving the linear programming problem. Let V_0 be the minimum value of the objective function given in (3.20) under the constraints (3.21) and (3.17). The value V_0 can be obtained by the simplex method using available computer programs.

Let,

$$LB2 \triangleq V_0/a \ .$$

Recalling the origin of the linear programming problem given by (3.20), (3.21) and (3.17) we have from (3.26) the following theorem.

THEOREM 3.4.2. *Under any service policy*

$$L \geq LB2 \ .$$

The following lemma follows directly from results appearing in several papers. (See, e.g., Baskett [2].)

LEMMA 3.4.1. *Under the service-sharing service policy, the stationary distribution of* η^t, $t \geq 0$ is

$$\mu(n) = \prod_{\alpha\epsilon A}\left\{(1 - \sum_{\beta\epsilon B} \rho_\alpha(\beta))\begin{pmatrix} n_\alpha(1)+n_\alpha(2)+,\ldots,+n_\alpha(b) \\ n_\alpha(1),n_\alpha(2),\ldots, n_\alpha(b)\end{pmatrix} \prod_{\beta\epsilon B} \rho_\alpha(\beta)^{n_\alpha(\beta)}\right\} ,$$

where $n = \{n_\alpha(\beta) | \alpha\epsilon A, \beta\epsilon B\}$.

The lemma will be used below, to derive another lower bound and an upper bound of the loss function.

THEOREM 3.4.3. *Under any service policy*

$$L \geq \sum_{\beta\epsilon B} (\min_{i\epsilon A} c_i(\beta)) \sum_{\alpha\epsilon A} \rho_\alpha(\beta)/(1-\rho_\alpha(\beta)) \triangleq LB3 .$$

Proof. For any $\beta_0\epsilon B$, consider the reduced systems $\Gamma^0_{\beta_0}$ obtained from Γ, consisting of customers of class β_0 only. (i.e., $B = \{\beta_0\}$.)

Let $\bar{n}^0_\alpha(\beta_0)$ be the expected number of customers of class β_0 at station α, under stationary conditions, in the system $\Gamma^0_{\beta_0}$.

It is clear that for any $\beta_0\epsilon B$ and under any given service policy in the system Γ,

$$\sum_{\alpha\epsilon A} \bar{n}_\alpha(\beta_0) \geq \sum_{\alpha\epsilon A} \bar{n}^0_\alpha(\beta_0) , \tag{3.28}$$

where $\bar{n}_\alpha(\beta_0)$ is the expected number of customers of class β_0 at station α under stationary conditions and under a given service policy in the original system Γ.

The values $\bar{n}^0_\alpha(\beta_0)$, are independent of the service policy taken in system $\Gamma^0_{\beta_0}$. Thus, it follows from lemma 3.4.1, when $B = \{\beta_0\}$, that

$$\sum_{\alpha\epsilon A} \bar{n}^0_\alpha(\beta_0) = \sum_{\alpha\epsilon A} \rho_\alpha(\beta_0)/(1-\rho_\alpha(\beta_0)) . \tag{3.29}$$

From (2.2), (3.28) and (3.29) we have

$$L = \sum_{\substack{\alpha\varepsilon A\\ \beta\varepsilon B}} c_\alpha(\beta)\overline{n}_\alpha(\beta) \geq \sum_{\beta\varepsilon B} (\min_{i\varepsilon A} c_i(\beta)) \sum_{\alpha\varepsilon A} \overline{n}_\alpha(\beta) \geq \sum_{\beta\varepsilon B} (\min_{i\varepsilon A} c_i(\beta)) \sum_{\alpha\varepsilon A} \rho_\alpha(\beta)/(1-\rho_\alpha(\beta)) \ .$$

REMARK 3.4.2.

We have LB2 $\geq$ LB1. However, LB3 is not always less than or greater than the bounds LB1, LB2. This can be seen in the examples appearing in Rosberg [17].

Let L_0 be the infimum of L over the set of service policies. Any value of L under any given service policy can be used as an upper bound for L_0. We shall use the Service-Sharing policy for that purpose, since in practice it is considered a good policy and the loss function under it can be computed.

From lemma 3.4.1 it follows that the value L under the Service-Sharing policy is

$$\sum_{\substack{\alpha\varepsilon A\\ \beta\varepsilon B}} c_\alpha(\beta)\rho_\alpha(\beta)/(1 - \sum_{j\varepsilon B} \rho_\alpha(j)) \triangleq UB \ . \tag{3.30}$$

Now, from (3.30) we have the following theorem.

THEOREM 3.4.4. $\quad UB \geq L_0$.

Theorems 3.4.1-3.4.4 are summarized in the following final corollary.

COROLLARY 3.4.1.

$$UB \geq L_0 \geq \max\{LB1, LB2, LB3\} \ .$$

4. DISCUSSION

In the model given in section 2.1 we assumed an exponential distribution of the service requirements. The reason behind this assumption is technical and lies in the fact that the distribution of the remaining service time of each customer when a transition occurs remains the same as at the moment of arrival. This is essential for an embedded Markov process analysis of the underlying process. These remaining time distributions cannot be handled in a network of queues with a general distribution of service requirements, since no matter what embedded Markov process we consider, there will always be a customer at an arbitrary point of his service duration when a transition occurs.

However, the same technique as in section 3, can be applied when the distributions of the service requirements are finite mixtures of gamma distributions. (In which the state space takes a much more complicated form.)

The idea and technique of using an embedded Markov process analysis with this generalization of the service requirements can be found in the literature (e.g., Baskett [2], Kelly [10] and Barbour [1]). The importance of this extension is that any nonnegative distribution can be approximated as closely as we please by a finite mixture of gamma distributions.

In this paper we considered only stationary service policies as defined in section 2.1 (i). An important problem which was not considered is under what conditions the optimal service policy (if it exists), among all possible service policies, is a stationary service policy. The known criteria for existence of an optimal stationary decision rule cannot be applied to the problem presented in this paper.

A question which should be answered before the question above concerns the conditions for ergodicity of the class of Markov processes η^t, $t \geq 0$, under any service policy. Some work on this last problem is done in Rosberg [16].

Finally, a use for the bounds and the structure of L presented in this paper, is given in a separate paper, Rosberg [17]. This paper presents an application of these results programmed for interaction computer use.

ACKNOWLEDGEMENT

The results appearing in this paper are part of the author's Ph.D. dissertation which was carried out in the Hebrew University of Jerusalem, under the supervision of Uri Yechiali and Shmuel Zamir to whom I am most grateful.

REFERENCES

1. Barbour, A. D., "Networks of Queues and the Method of Stages," *Adv. Appl. Prob.* 8, 3, 584-591 (1976).

2. Baskett, R., Chandy, K. M., Muntz, R. R. and Palacios, F. G., "Open, Close and Mixed Network of Queues with Different Classes of Customers," *J. ACM.* 22, 2, 248-260 (1975).

3. Chandy, K. M., Howard, J. H. Jr. and Towsley, D. F., "Product Form and Local Balance in Queueing Networks," *J. ACM.* 24, 2, 250-263 (1977).

4. Cinlar, E., *Introduction to Stochastic Processes,* Prentice-Hall, Englewood Cliffs, New Jersey, 1975.

5. Cox, D. R. and Smith, W., *Queues,* Methuen, London, 1961.

6. Harrison, J. M., "Dynamic Scheduling of a Multiclass Queue: Discount Optimality," *Opns. Res.,* 23, 2, 270-282 (1975).

7. Jackson, J. R., "Networks of Waiting Lines," *Opns. Res.* 5, 518-521 (1957).

8. Jansen, U., König, D. and Nawrotzki, K., "A Criterion of Insensitivity for a Class of Queueing Systems with Random Marked Point Processes," Bergakademie Freiberg, Sektion Mathematik, DDR 92 Freiberg (1978).

9. Kakalik, J. S., "Optimal Dynamic Operating Policies for a Service Facility," Technical Report No. 47, Operations Research Center, M.I.T. (1969).

10. Kelly, F. P., "Networks of Queues," *Adv. Appl. Prob.* 8, 2, 416-432 (1976).

11. Kleinrock, L., *Queueing Systems Vol. II: Computer Applications,* John Wiley & Sons, New York 1976 .

12. Klimov, G. P., "Time-Sharing Service Systems, I," *Theory of Prob. and Its Appl.* 19, 3, 532-551 (1974) (English Translation).

13. Lippman, S. A., "Applying a New Device in the Optimization of Exponential Queueing Systems," *Opns. Res.* 23, 4, 687-710 (1975).

14. Melamed, B., "On Poisson Traffic Processes in Discrete-State Markovian Systems with Application to Queueing Theory," *Adv. Appl. Prob.* 11, 1, 218-239 (1979).

15. Noetzel, A. S., "A Generalized Queueing Discipline for Product Form Network," *J. ACM.* 26, 4, 779-793 (1979).

16. Rosberg, Z., "A Positive Recurrence Criterion Associated with Multi-Dimensional Queueing Processes," *J. Appl. Prob.*, September (1980) (To appear).

17. Rosberg, Z., "Queueing Networks under the Class of Stationary Service Policies, II," Working Paper No. 647, College of Commerce and Business Administration, University of Illinois at Urbana-Champaign, February 1980.

18. Schassberger, R., "The Insensitivity of Stationary Probabilities in Networks of Queues," *Adv. Appl. Prob.* 10, 906-912 (1978).

19. Schrage, L. E., "A Proof of the Optimality of the Shortest Remaining Processing Time Discipline," *Opns. Res.* 16, 687-690 (1968).

BOUND BY
HW
3-94

UNIVERSITY OF ILLINOIS-URBANA
3 0112 060296305